DOCUMENT RESUME

ED 360 361 TM 020 263

AUTHOR Bethscheider, Janine K.
TITLE Internal-Structure Analysis of Analytical Reasoning Worksamples 244 D and E and Development of Form H. Technical Report 1992-1.
INSTITUTION Johnson O'Connor Research Foundation, Chicago, IL. Human Engineering Lab.
PUB DATE Dec 92
NOTE 42p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adults; *Aptitude Tests; Career Choice; Comparative Testing; Correlation; Difficulty Level; Educational Objectives; *Item Analysis; *Test Construction; Test Format; Test Interpretation; Test Items; Test Reliability; *Thinking Skills
IDENTIFIERS *Analytical Reasoning Test (Johnson O Connor); *Internal Structure Analysis; Johnson O Connor Aptitude Tests

ABSTRACT

Standard and experimental forms of the Johnson O'Connor Research Foundations Analytical Reasoning test were administered to 1,496 clients of the Foundation (persons seeking information about aptitude for educational and career decisions). The objectives were to develop a new form of the test and to better understand what makes some items more effective than others. Internal-structure analysis of the 26 new items indicated that all but 3 could be regarded as at least adequate on the basis of their correlation with current standard forms of the test. Thirteen items with the highest item-total correlations or the greatest contribution to overall test reliability were selected for the new test version. The relationships of six item characteristics to item quality and item difficulty were investigated, and these suggest that the clarity of the conceptual features among the words or concepts contributes more to item quality than any of the features studied. Suggestions are made for developing a new alternative form of the Analytical Reasoning test, drawing on the items not included on this revised version. Two tables present study data, and three figures illustrate the discussion. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERT KYLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED360361

INTERNAL-STRUCTURE ANALYSIS OF ANALYTICAL REASONING WORKSAMPLES 244 D AND E AND DEVELOPMENT OF FORM H

Janine K. Bethscheider

JOHNSON O'CONNOR RESEARCH FOUNDATION, INC.

Technical Report 1992-1

December 1992

BEST COPY AVAILABLE

CE 062263

COPYRIGHT © 1992 BY JOHNSON O'CONNOR RESEARCH FOUNDATION, INCORPORATED
ALL RIGHTS RESERVED

BEST COPY AVAILABLE

Internal-Structure Analysis of Analytical Reasoning Worksamples 244 D and E and Development of Form H

Janine K. Bethscheider

ABSTRACT

Standard and experimental forms of the Foundation's Analytical Reasoning test were administered to 1,496 clients of the Johnson O'Connor Research Foundation. The primary objectives of this research project were to develop a new form of Analytical Reasoning with improved reliability and obtain a better understanding of what differentiates more-effective Analytical Reasoning items from less-effective ones.

Internal-structure analysis of the 26 items indicated that all but three can be regarded as at least adequate on the basis of their correlations with the current standard form of the test. In order to maximize reliability, the items that had the highest item-total correlations or contributed most to overall test reliability were selected initially for the pool of items for the new version of the test, and estimates of alpha reliability were then assessed for various combinations of the best items. A set of 13 items with near-maximal reliability that makes use of the existing Analytical Reasoning playing boards was selected for the next standard form of the test. The items for this new form were then arranged in approximate order of difficulty, and a system of scoring using the traditional stanine method was chosen. The estimated alpha reliability of this newly developed 13-item test is .827 for the summed stanine scores when the effects of age and sex are removed.

The relationships of six item characteristics including asymmetry to item quality and item difficulty were also investigated. The findings suggest that, in general, the fewer the number of chips involved, the easier the item. More important, no systematic relationships of item characteristics to item quality were identified. Some of the Analytical Reasoning items were more effective than others, independent of their configurational features. In all likelihood, clarity of the conceptual structures among the words or concepts contributes more to item quality than any of the item features studied.

In light of these findings, the suggestion is made that any further development of test items focus on making the relationships among the words or concepts clear. In addition, it is recommended that future Foundation research be directed towards identifying acceptable items for an alternate form of the Analytical Reasoning test rather than refining further the new form. To determine whether a combination of items from the standard and experimental forms that were not included on the new form might adequately serve as an alternate test, the psychometric properties of a test composed of 11 of the 13 leftover items were assessed. It was concluded that it might be possible to construct an alternate form with close-to-acceptable reliability by using seven of these leftover items and modifying or replacing the other six items.

In summary, a new, more-reliable form of the Foundation's Analytical Reasoning test was developed. With further research, it might also be possible to construct an alternate form of the test by using seven items not included on the new form and developing six additional items.

CONTENTS

	Page
Introduction	1
Method	1
Examinees	1
Measures	2
Administration and Scoring	4
Analyses	7
Results and Discussion	7
Distribution of Scores	7
Internal-Structure Analysis of Standard and Experimental Items	8
Selection of Items and Development of a More-Reliable Worksample	10
Relationships of Item Characteristics to Item Quality and Item Difficulty	17
Correlations and Factor Analysis	21
Sex and Age Effects	22
Additional Analyses	23
General Summary and Conclusions	28
References	32

LIST OF TABLES

		Page
Table 1	Item Analysis for Analytical Reasoning Worksamples 244 D and E	9
Table 2	Difficulties of Items on Analytical Reasoning Worksample 244 H	15

LIST OF FIGURES

		Page
Figure 1	Sample Playing Board for Item from Analytical Reasoning	3
Figure 2	Examples of Right-Side-Up and Upside-Down Solution for an Analytical-Reasoning Item	5
Figure 3	Age Curve for Analytical Reasoning Worksample 244 H	24

ACKNOWLEDGMENTS

The author would like to express her appreciation to the following persons for their contributions to this study: Robert Kyle, Director of Research, who assembled the experimental test of Analytical Reasoning for this project; the test administrators and directors of the Foundation's testing offices, who collected the data for this project; and the support staff of the Research Department, who assisted in the preparation of the data for analysis. In addition, the author gratefully acknowledges David Schroeder, who as Research Manager made invaluable recommendations with regard to this report.

INTRODUCTION

One concern of the Foundation with regard to the Analytical Reasoning test has been its lower-than-desirable reliability. The reported internal-consistency reliability for the 13-item test (Form CM) is .65 (Statistical Bulletin 1974-13), which means that scores on Analytical Reasoning contain a relatively large proportion of error (35%). (See Statistical Bulletin 1988-2 for a thorough discussion of reliabilities and standard errors of measurement.)

One way to improve a test's reliability is to lengthen the test by adding items. Another method is to replace items that detract from or contribute little to the test's reliability with items that enhance its reliability. In view of this, in 1989, an experimental form of Analytical Reasoning, Form E, was assembled and administered to Foundation examinees in an effort to identify a pool of reliable items that might serve as replacements or additions to the standard test items. The objective of this project was to create a more reliable form of the Analytical Reasoning test.

Recently, two items were removed from the standard Analytical Reasoning test (Test Information Bulletin 1991-9), reducing Analytical Reasoning from 13 items (Form D) to an 11-item test (Form F). This has created an even greater need for additional items that will contribute positively to the internal consistency of the test.

This report presents the results of an internal-structure analysis of standard and experimental items for Worksample 244, Forms D and E, including item analysis and internal consistency estimates of reliability. In addition, it discusses the development of a new form of Analytical Reasoning with improved reliability.

METHOD

Examinees

The examinees for this study were paying clients of the Johnson O'Connor Research Foundation (JOCRF) who were tested in seven of the Foundation offices across the United States. These examinees came to the Foundation for testing in order to obtain information about their aptitudes that they could use in making educational and occupational decisions. The Foundation's examinee population is a relatively homogeneous group with respect to education and socioeconomic status. Foundation clients typically are white and middle- to upper-middle-class. The majority are college-bound or college-educated.

A total of 1,496 Foundation clients completed both the standard and experimental versions of the Analytical Reasoning test. Of these, 782 were male (52.3%) and 714 were female (47.7%). The age of the participants ranged from 14 to 65, with a mean of 28.7 (SD = 10.1) and a median of 26. Approximately 31.8% of the examinees were tested in Foundation offices located in the southern United States (Dallas, New Orleans, Tampa), 30.0% in the East (New York), 19.4% in the West (Los Angeles, Denver), and 18.8% in the Midwest (Chicago).

Measures

The Analytical Reasoning test that was administered to Foundation examinees in 1989 as part of the standard battery was Form D, which consists of a practice item and 13 test items. Form E, the experimental version of Analytical Reasoning that was administered as part of this project, is patterned after Form D, except that a practice item is not administered with Form E. That is to say, identical item diagrams, or board configurations, were used for both worksamples, and these configurations are presented in the same order on both tests. This means that Item 1 of Form E uses the same item diagram as Item 1 of Form D. Similarly, Items 2 through 13 of Form E have the same configuration as Items 2 through 13 of Form D, respectively.

The items included on Form E were selected from tests of Analytical Reasoning previously developed within the Foundation. Specifically, Items 1 through 11 of Wks. 244E were chosen from among the 40 items that make up Worksamples 696A and 696B, group-administered Analytical Reasoning tests that were constructed in 1980. Kyle (Statistical Bulletin 1989-2) reported that "an attempt was made to choose items showing the highest correlation with Wks. 244C [the standard form of Analytical Reasoning in 1980] and the lowest correlation with English vocabulary. Due to a lack of items of certain patterns, Items 12 and 13 were selected from an experimental version of Analytical Reasoning created by Jennifer Osborn in 1974" (p. 1).

As noted earlier, Form F was introduced as the standard Analytical Reasoning test in the spring of 1991 (i.e., after the data collection phase of this project was concluded) and comprises Items 1-10 and 12 of Form D; in all other respects, Forms D and F are identical. Items 11 and 13 were dropped from the standard test because the "solution of these two items depends on acquired knowledge in addition to reasoning ability" (Test Information Bulletin 1991-9, p. 1). These two deleted items ("Country" and "Clothes") are also known as the "silkworm" items because both use the word *silkworm*.

In each item of the Analytical Reasoning test, the examinee is presented with a playing board with hexagonal spaces printed across the top and a design, or diagram, of blank hexagons interconnected by arrows in the center. Hexagonal chips with words printed on them are placed in alphabetical order in the spaces across the top of the board, and the examinee is asked to arrange these word-chips on the blank spaces in the middle of the board in such a way that they form a logical sequence. Figure 1 presents a sample playing board (using squares rather than hexagons) for a test item from Forms D and E.

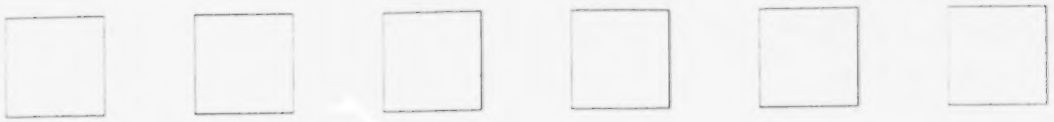
Forms D and E of the Analytical Reasoning test utilize seven different board configurations, which can be described in terms of various combinations of certain item features, the most salient of which are the three mentioned below.

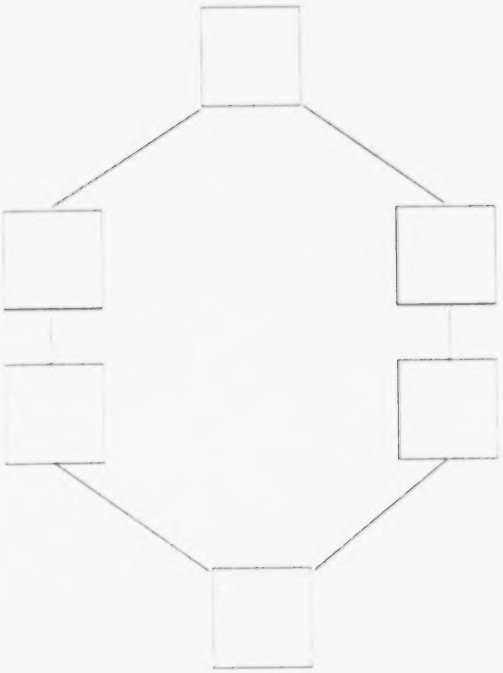
1. Number of word-chips involved. Of the seven types of configurations used in Forms D and E, one is a 3-chip diagram, two are 5-chip designs, three are 6-chip designs, and one involves a 7-chip diagram. Figure 1 depicts a 6-chip diagram.

2. Symmetry or asymmetry of the item configuration. Only one asymmetric diagram type is used in Forms D and E; the other six item configurations are symmetric about a

Figure 1

Sample Playing Board for Item from Analytical Reasoning





vertical line through the center of the playing board. An example of a symmetric diagram is shown in Figure 1.

3. "Circular" versus "middle-converging" versus "branched" configuration. Circular diagrams are those that begin with one chip and end with one chip, such as those for Items 2 and 9 (on both Forms D and E). Middle-converging configurations have one or two chips in the center of the diagram, such as those for Items 3 and 13. Branched diagrams begin with one chip and then branch out to end with more than one chip, such as those for Items 1 and 7. Three of the seven diagram types used in Forms D and E are circular, two are middle-converging, and two are branched. Figure 1 is an example of a circular configuration.

In terms of these three features, then, the seven board configurations used with Forms D and E of the Analytical Reasoning test can be characterized as follows (the number in parentheses indicates the number of items on Forms D and E of that particular configuration):

1. 3-chip symmetric branched (1)
2. 5-chip symmetric branched (1)
3. 5-chip symmetric circular (1)
4. 6-chip symmetric circular (4)
5. 6-chip symmetric middle-converging (1)
6. another 6-chip symmetric middle-converging design distinct from Diagram 5 (2)
7. 7-chip asymmetric circular (3)

In addition, for some of the diagrams the word-chips can be arranged in such a way that they would form a correct sequence if the directions of all the arrows in the diagram were reversed and the diagram were turned upside down (i.e., rotated 180°). In such an arrangement of word-chips, which the Foundation refers to as an upside-down solution, the chip that would be at the top of the diagram in the right-side-up solution is placed instead on the bottom of the diagram, the chip that normally would be on the bottom of the diagram is placed at the top, and so on. (Figure 2 provides an example of a right-side-up solution as well as a corresponding upside-down solution.) Upside-down, as well as right-side-up, solutions are possible with three of the seven board configurations used in Forms D and E, namely, Diagrams 4, 5, and 7.

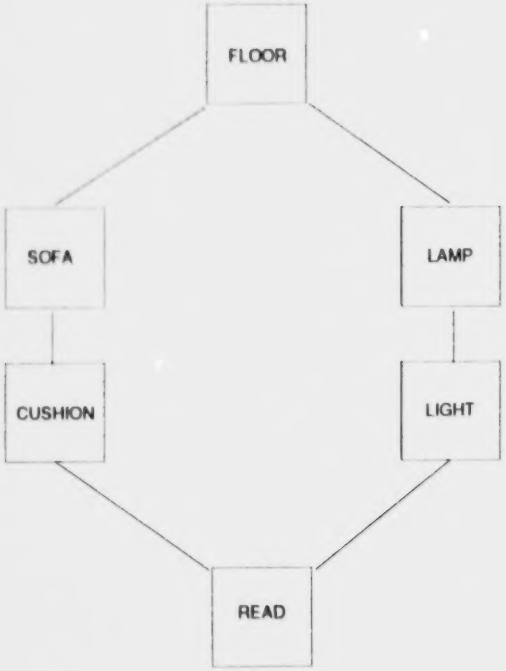
Administration and Scoring

The worksamples were individually administered, one item at a time. Form D was administered as part of the standard battery during the first individually administered session of testing; Form E was given at the end of the second individual session. Oral instructions preceded the tests. The examinee was told to work as quickly as possible, beginning as soon as the administrator placed the last chip down on the board.

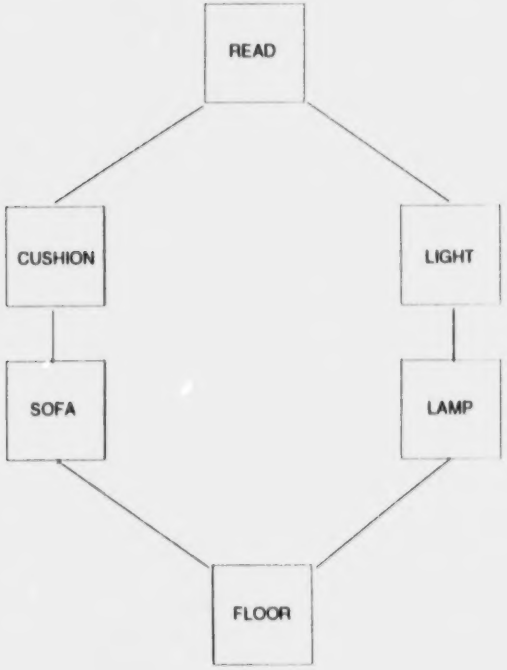
If the examinee incorrectly arranged the word-chips, the administrator responded "No, keep trying" (*Scoring Guide*, 1987, p. 57). If an item on Form D was not completed correctly within a certain time limit, coaching by the administrator was permissible. For examinees who were doing poorly on the standard test, administrators were permitted to terminate the test early following rules in the *Scoring Guide* (1987). For the experimental test, however, examinees were required to complete all 13 items.

Figure 2

Examples of Right-Side-Up and Upside-Down Solution for an Analytical-Reasoning Item



(a) Right-side-up solution



(b) Upside-down solution

For this study, an examinee's score for each item was the time it took the examinee to arrange the word-chips correctly. Upside-down, as well as right-side-up, solutions were counted as correct. No penalty was assessed for incorrect arrangements of the word-chips. For each item, there was a cutoff time beyond which an examinee could be stopped or coached on the item. Consequently, if an examinee failed to solve a given item within this specified time, he or she was assigned the score equal to the cutoff time for that item. For Form D, the cutoff time for Items 1 through 7 was 1.0 minutes; for Items 8 through 13, the cutoff time ranged between 1.18 and 2.01 minutes, depending on the individual item. For each item on Form E, the cutoff time was 1.5 minutes. Because the items on the standard test were arranged in order of difficulty, it was assumed that examinees who could not solve the easier items in less than their cutoff times would not be able to solve the more-difficult items in less than their cutoff times. Consequently, if Form D was terminated before all 13 items were administered, the score for the cutoff time was assigned to each item not taken. If an individual time was marked "not for research" (because, for example, the timer malfunctioned), it was considered a "missing" value.

An examinee's score for the full standard worksample was calculated by adding together the examinee's 13 item scores on Form D. Similarly, an examinee's score for the full experimental worksample was computed by totaling the examinee's 13 item scores on Form E. If an examinee was missing an individual item score, no total score was calculated for the form on which the item appeared.

When the scoring of a test is based on time to solution rather than number completed correctly within a given amount of time, a nonlinear transformation of the time-based scores is psychometrically appropriate in order to obtain scores that more closely approximate interval-level measurement (Cohen & Cohen, 1983, pp. 253-255). Before the transformed scoring method for the Analytical Reasoning test could be chosen, however, several analyses had to be performed.

In brief, the test items were scored using a number of scoring systems, including square roots of the item times, reciprocals of the item times, reciprocals of the square roots of the item times, and logarithms of the item times. Then, the shapes of the individual item-score distributions that resulted from the various transformation methods were determined. In addition, an internal-consistency reliability coefficient was calculated for each of these methods. The resulting distributions and reliabilities served as the bases for assessing the various systems. Specifically, each scoring method was evaluated on two criteria:

1. Did it result in more-normally distributed item scores (and not just total test scores)?

2. Did it maximize reliability?

For this project, the transformation that best satisfied these criteria was taking the logarithm of the time for each item. (The alpha coefficient for Form D using logarithms of the item times was .76, compared with an alpha coefficient of .68 using untransformed item times and .75 using the Foundation's one-to-four-point scoring system.) An examinee's transformed total score for the standard worksample, then, was the sum of his or her transformed item scores on Form D (i.e., the sum of the logarithms of the item times rather than the logarithm of the sum of the item times). Likewise, an examinee's transformed total score for the experimental worksample was the sum of the examinee's transformed item scores on Form E.

Analyses

The principal objectives of this research project were to develop a form of Analytical Reasoning with improved reliability and obtain a better understanding of what differentiates the more-effective Analytical Reasoning items from the less-effective ones in order to make recommendations regarding the composition of future items. Tô address these objectives, several series of analyses were performed on the data collected in this study. These analyses are discussed briefly below.

The initial analyses examined the distribution of total test times along with the internal structure of the standard and experimental items, which included item analysis and internal-consistency estimates of reliability. The next series of analyses dealt with the development of a more-reliable form of the Analytical Reasoning test. Based on the results of the item analysis, the Analytical Reasoning items were evaluated individually in terms of their item-total correlations and unique contributions to overall test reliability, and various combinations of the better items were then assessed. Of primary interest were those combinations that maximized test reliability. After the selection of a set of items for the new form of Analytical Reasoning, the presentation order of the items was established, and a scoring system for the new test was defined. The third set of analyses explored the relationships of various item features to item quality and, to a lesser extent, item difficulty.

Besides the aforementioned series of analyses, several additional analyses were performed on the data. Correlational analyses and factor analyses were performed on the standard and experimental item scores to investigate the relationships among the 26 items and determine whether the items are measuring a single underlying trait. Sex and age differences in analytical-reasoning ability were assessed using the regression approach to analysis of variance of test scores by sex and age. Other analyses included the following: additional reliability analyses for the proposed new form as well as Forms D and F; comparisons of Form D with Form CM and the new form; evaluation of the relationship between the current standard test and the new form; and investigation of the relationships between sex and age and performance on the "silkworm" items.

The results of the analyses for this study are reported in the following sections. Unless otherwise stated, the *SPSS/PC+* (Version 4.0; Norusis, 1990a) and *SPSS/PC+ Statistics* (Version 4.0; Norusis, 1990b) computer software packages were used for the analyses.

RESULTS AND DISCUSSION

Distribution of Scores

Total test times for the 13-item test that was standard at the time of the study (i.e., Form D) ranged from 2.4 minutes to 13.3 minutes (M = 7.44, SD = 1.95). Total test times for the 11 items that now comprise the standard test (i.e., Form F) ranged from 1.2 to 9.9 minutes (M = 5.43, SD = 1.56). For both of these forms, the distribution of total test times approximated the normal curve. There were no appreciable floor or ceiling effects, and total times were spread widely enough across the range for differences among examinees to exhibit themselves.

For the experimental test (Form E), total test times ranged from 1.0 to 17.2 minutes (M = 4.39, SD = 1.99). The distribution was somewhat skewed, with times concentrated toward the low end of the range. When the times were transformed to form scores (see the Administration and Scoring section), however, the resulting distribution of total times for the experimental form as well as the 11- and 13-item standard forms approximated the normal curve. Therefore, transformed scores were utilized for most of the analyses for this project.

As a group, the experimental items were easier for examinees than the items on the standard test. Over 90% of the examinees completed Form E within 7 minutes, compared with 42% who finished Form D in that amount of time. Within a 10-minute time period, 98% had completed Form E, and 90% Form D. It should be noted, however, that because Form E was always administered after Form D, part of this effect may be attributable to practice.

Internal-Structure Analysis of Standard and Experimental Items

Item Difficulties

Values of the item statistics for the Analytical Reasoning tests are displayed in Table 1. The first four columns present the untransformed and transformed means and standard deviations for the standard and experimental items. The item scores are based on time to solution, with lower values indicating easier items.

The average solution time for the 13 original items was 60.6 hundredths of minutes (SD across items = 38.5) and, with the "silkworm" items deleted, 51.1 (SD across items = 31.7). For the experimental form, the mean and standard deviation were 47.2 and 33.1, respectively.

As can be observed from Table 1, the mean time to solution ranged from 6.9 to 138.5 hundredths of minutes on Form D. Item 1 (the "Roast" item) showed the fastest time to completion. The "Cigarette" item manifested the slowest time to completion on Form D, except for one of the "silkworm" items ("Clothes"). On Form E, mean time to solution ranged from 3.5 hundredths of minutes for the "Heat" item to 73.6 for the "Cloth" item.

The percent of examinees who solved each item in less than a minute also is contained in Table 1, as an alternative gauge of item difficulty. All the examinees were able to solve the two three-chip items ("Roast" and "Heat") within a minute. In addition, three of the experimental items ("Breakfast," "Chair," and "Eat") were completed correctly within this time period by over 99% of the examinees. Overall, six of the standard items and nine of the experimental items were solved in less than a minute by at least 90% of the participants. Again, one of the "silkworm" items ("Clothes") proved to be the most difficult to solve, with only one-third of the examinees completing it correctly in less than a minute.

Item-total Correlations

Table 1 also displays the item-total correlations for the standard and experimental items. In the calculation of the item-total correlations, examinees' transformed scores on each item were correlated with their transformed total test scores on Form F (not the total score for the item's form). To adjust for part-whole overlap, the corrected item-total correlation for each item on Form F was determined by correlating the item score with the total score minus the

Table 1

Item Analysis for Analytical Reasoning Worksamples 244 D and E

Item #	1st chip	Untransformed Mean solution time	Untransformed *SD*	Transformed Mean solution time	Transformed *SD*	Percent with time < 1.00	Item-total corr.[a]
Standard test (Form D):							
1	ROAST	6.85	6.52	1.66	.69	100.0	.40
2	COW	24.45	17.28	3.03	.55	98.4	.48
3	HUMAN BEING	25.88	20.18	3.04	.61	97.3	.41
4	FISH	34.41	26.07	3.30	.68	93.2	.43
5	CAKE	31.75	24.23	3.23	.66	94.3	.43
6	AIR	38.98	24.81	3.48	.61	94.0	.43
7	AUTO. FACTORY	56.55	33.95	3.79	.78	73.8	.29
8	ATTACK	63.34	38.62	3.92	.71	71.8	.24
9	BROOM	82.84	48.04	4.22	.66	63.5	.35
10	CUSHION	87.59	49.81	4.27	.68	60.2	.29
11	COUNTRY[b]	87.04	48.38	4.27	.66	58.8	.46
12	CIGARETTE	109.38	70.78	4.43	.78	53.3	.37
13	CLOTHES[b]	138.54	62.86	4.77	.65	33.6	.32
Experimental test (Form E):							
1	HEAT	3.53	3.28	1.09	.55	100.0	.37
2	ARROW	28.93	27.27	3.09	.69	96.0	.31
3	COOK	34.42	30.53	3.28	.67	94.0	.43
4	BREAKFAST	14.29	9.31	2.54	.45	99.9	.51
5	APPLAUSE	21.75	21.12	2.85	.61	98.0	.43
6	ASTRONAUT	51.15	44.09	3.59	.82	83.1	.32
7	CHAIR	15.87	15.20	2.55	.59	99.1	.41
8	EAT	13.84	14.26	2.42	.56	99.2	.36
9	BREAD	40.99	34.82	3.46	.67	91.4	.45
10	BAG	34.42	29.31	3.28	.68	94.9	.42
11	CLOTH	73.64	49.68	4.04	.76	68.6	.37
12	FURNITURE	45.63	39.88	3.52	.74	87.4	.43
13	ARTIST	60.36	44.72	3.84	.73	79.1	.41

[a]For the item-total correlations, the total test score used in the calculations is the total score for Form F, the 11-item current standard Analytical Reasoning worksample (not the total score for the item's form). The item-total correlations are corrected for overlap between the individual items and the score for the 11 Form F items.

[b]Item contains the word-chip "silkworm."

item score. For the experimental and "silkworm" items, no correction for item-total overlap was necessary; that is, their item-total correlations were determined by correlating each item score with the total score for Form F.

Item-total correlations for the 26 items ranged from .24 to .51. Three of the standard items ("Attack," "Auto Factory," and "Cushion") displayed relatively low item-total correlations. The remaining 23 standard and experimental items, including the "silkworm" items, correlated at least .31 with the score for the 11 Form F items. In other words, on the basis of item-total correlations, all but three of the Analytical Reasoning items can be regarded as at least adequate.

Internal Consistency

Coefficient alpha was computed as the estimate of reliability for the standard and experimental forms of the Analytical Reasoning test. When the transformed scores for the 11 items that comprise Form F were used in the calculation, an alpha of .724 resulted. When all 13 standard items (Form D) were included in the total score, a reliability of .760 was obtained. All the standard items contributed positively to the internal consistency of the overall test, although the "Attack" item's contribution was negligible; that is, if the item were deleted, Form D's reliability would be reduced by only .0003 and Form F's by .0007. The alpha coefficient for Form E was .834. All 13 items made a positive contribution to the reliability of the experimental worksample, although deleting the "Astronaut" item from Form E would reduce the test's reliability by less than .002.

When all 26 items were included in the total score, an alpha of .880 resulted. When all but the "silkworm" items were used in the calculation, a reliability of .873 was obtained. The "Attack" item made a very slight negative contribution to the internal consistency; that is, deleting the item would increase overall test reliability by .0001. The "Auto Factory" item's contribution to test reliability also was negligible; that is, if the item were deleted, the reliability of the 26-item test would be reduced by only .0001, and the reliability of the 24-item test would remain unchanged.

Selection of Items and Development of a More-Reliable Worksample

Item Selection

In order to determine which of the standard and experimental items should be selected for the new, more-reliable version of the Analytical Reasoning test, several indices of item quality were employed. Each item was evaluated in terms of its (a) corrected item-total correlation and (b) unique contribution to the overall reliability of the test. Then the alpha reliabilities for various combinations of the better items were calculated and compared. These analyses were performed using transformed time scores.

Initially, no restrictions were placed on the number or type of items that might be included in the new test, and the following selection criteria were used to form the various item combinations:

1. The items were ranked on the basis of their item-total correlations, and those items with the highest item-total correlations were selected.

2. Each item's unique contribution to the test's internal consistency was determined, and the items that contributed least to alpha were removed from the item pool.

Because the two "silkworm" items had just recently been deleted from the standard test and it was unlikely that they would be reinstated (at least without any revision), only the results from the analyses that excluded these two items will be discussed in this section, unless otherwise stated. (All the selection procedures were in fact performed twice: once with the "silkworm" items included at the beginning of the selection process and once with them excluded from the start.) In addition, because of the similarity between the "Cow" item on Form D and the "Breakfast" item on Form E, it was decided that only one of these items should be included in the analyses, namely the "Cow" item. The reason for choosing the "Cow" item was that the item statistics for the "Breakfast" item were of questionable accuracy because examinees had already taken a nearly identical item, the "Cow" item. Thus, we know that the "Cow" item is an adequate item, whereas we are not certain about the "Breakfast" item.[1]

Obtained reliability estimates for some of the initial item combinations were as follows. When the three items with corrected item-total correlations below the level of acceptability ("Auto Factory," "Attack," and "Cushion")--the three items that also contributed the least to test reliability--were deleted from the total score, the alpha reliability of the test declined by .0003 (i.e., alpha equaled .864). When the test was shortened to the best 15 items (based on item-total correlations), an alpha of .841 was obtained; the best 14 items, .837; and the best 13 items, .829. For the test composed of the best 15 items based on unique contributions to reliability, an alpha of .845 was obtained; for the corresponding 14-item test, .839; and for the corresponding 13-item test, .832. (Inclusion or exclusion of one of the "silkworm" items ["Clothes"] had little effect on these analyses because it had the fifth lowest item-total correlation and was the fifth item to be removed from the item pool on the basis of its contribution to alpha. The other "silkworm" item ["Country"] had the second highest item-total correlation and was the sixteenth item to be removed from the item pool on the basis of its unique contribution to test reliability. Consequently, if the "Country" item had been included in the pool of possible items for the new test, it would have been selected as one of the better items on the basis of item-total correlations and unique contributions to reliability; nevertheless, the reliability estimates reported above would have been raised by no more than .0024.)

Based on the results of these analyses, the following decisions were made:

1. To restrict the number of items that could be selected for the new test to no more than 13 to 15. In this way, the new test would not take much longer to administer than the original 13-item standard test, and yet the sacrifice to reliability would be .036 at most

---

[1]Omitting the "Breakfast" item from the calculations of internal consistency resulted in an alpha of .874 when all 25 test items were included in the total score and .865 when all but the "silkworm" items were included. The "Attack" and "Auto Factory" items' contributions to test reliability were negligible. That is to say, if the "Attack" item were deleted, the reliability of the 25- and 23-item tests would be reduced by only .0001; if the "Auto Factory" item were deleted, the reliability of the tests would be reduced by .0002.

(relative to the full 23-item set), while alpha would remain in an acceptable range (i.e., above .80).

2. To include a 3-chip item on the test, even if it would otherwise be excluded on the basis of the selection criteria. The inclusion of this condition did not affect test reliability, however, because at least one of the 3-chip items was already included in the 13-, 14-, and 15-item tests consisting of the best items, whether based on item-total correlations or unique contributions to reliability.

At this point in the process of developing a new Analytical Reasoning test, a third selection criterion was introduced. Whereas the first two criteria focused on maximizing reliability without any restrictions, the third criterion was directed towards maximizing reliability within a given restriction. Specifically, the restriction was that all 13 playing boards used with Form D also be utilized with the new form. The rationale for considering this possibility was that it would make the transition to the new form easier.

In order to maximize reliability within this restriction, the standard and experimental items were grouped according to diagram type and then, within each group, the items that had the highest item-total correlations and/or contributed most to overall test reliability were selected for the item pool. The proviso was that the board configurations of 13 of the selected items had to correspond exactly (in terms of number and type of diagram) to the playing boards that were used with Form D. For instance, because three 7-chip asymmetric and circular items were included on Form D, three such items had to be included in the item pool. Likewise, because one of Form D's playing boards was diagrammed with a 5-chip symmetric, branched configuration, all item combinations selected on the basis of this third criterion had to include one item with that configuration.

Furthermore, for the 14- and 15-item combinations, a second restriction was imposed, namely, that the additional one or two items would be selected from among the diagram types represented by three or fewer items on Form D. The purpose of this condition was to maintain a balance of the types of items selected, so that none of the seven board configurations would be overrepresented on the new test. (In practice, no more than four items with the same board configuration were included in the item pool; in practical terms, this meant that only four of the eight identically diagrammed 6-chip symmetric and circular items could be selected.) Thus, by selecting items on the basis of the third criterion, no more than 31% of the items on a 13-item test would share the same board configuration, 29% on a 14-item test, and 27% on a 15-item test. (By way of comparison, selecting items solely on the basis of their item-total correlations or unique contributions to reliability could, in theory at least, have resulted in a 13-item test with up to 62% of the items identically configured, up to 57% in a 14-item test, and up to 53% in a 15-item test.)

The effects of these restrictions on alpha were then explored. For tests composed of the 15 best items (based on maintaining a balance of item configurations), reliability estimates ranged from .841 to .842; for the 14-item tests, from .836 to .838; and for the 13-item tests, from .829 to .832. Thus, it appears that, without sacrificing reliability, item combinations can be formed that (a) use the 13 playing boards of Form D and at the same time (b) maintain a balance of item types for the 14- and 15-item combinations.

From all the item combinations for which reliability coefficients had been calculated, six were then chosen for further investigation: the best 13-, 14-, and 15-item sets (based on unique contributions to reliability) and the best 13-, 14-, and 15-item sets (based on maintaining a balance of item configurations). Inspection of the composition of these six item sets led to the identification of a core group of 12 items--items that were included in all six of the best item combinations: "Cow," "Fish," and "Air" from Form D and "Heat," "Cook," "Applause," "Chair," "Eat," "Bread," "Bag," "Furniture," and "Artist" from Form E. In addition, the "Human Being" item was included in all four of the 14- and 15-item combinations.

The only difference, then, between the best 13-item test (based on unique contributions to reliability) and the best 13-item test (based on maintaining a balance of item configurations) was that the former (alpha = .832) included the "Roast" item from the standard worksample and the latter (alpha = .832) included the "Cloth" item from the experimental worksample. Likewise, the only difference between the best 14-item test (based on unique contributions to reliability) and the best 14-item test (based on maintaining a balance of item types) was that the former (alpha = .839) included the "Roast" item and the latter (alpha = .838) included the "Cloth" item. The differences between the best 15-item test (based on unique contributions to reliability) and the best 15-item test (based on maintaining a balance of items) were that the former (alpha = .845) included the "Roast" and "Cake" items from the standard worksample whereas the latter (alpha = .842) included the "Broom" item from the standard worksample and the "Cloth" item from the experimental worksample.

One additional comment is in order here. The 13-, 14-, and 15-item tests with the best reliability consisted of items selected on the basis of their unique contributions to reliability and included only two asymmetric 7-chip items (i.e., Diagram 7). In order to utilize all 13 playing boards from Form D, a third item with that configuration (specifically, the "Cloth" item) had to be inserted into the item pool. Nevertheless, substituting the "Cloth" item for a differently diagrammed item that contributed more to overall test reliability (specifically, the "Roast" item) lowered alpha by only approximately .003 or less.

Based on these findings, several conclusions were reached. The first was that the 13 playing boards from Form D should in fact be utilized with the new test. Using the same boards will have a minimal effect on the internal consistency of the new test and will necessitate the manufacture of, at most, two additional boards. The second conclusion was that either a 13- or 14-item test should be developed rather than a 15-item test, given that all three are acceptable in terms of overall test reliability. The reasoning behind this was that the current box that holds the word-chips has 14 slots, which means it could hold the chips for a practice item and 13 test items or, if the chips from the practice item and the 3-chip test item were combined in one slot, it could accommodate a 14-item test. In order to implement a 15-item test, however, either new boxes would have to be produced or the two 3-chip groups would have to be stored somewhere else. Because (a) both the best 13- and 14-item tests (based on maintaining a balance of items) have acceptable alpha reliabilities and (b) the difference in their reliability estimates is minimal, the decision was made to construct the new Analytical Reasoning worksample from the most reliable of the 13-item combinations that utilized the 13 playing boards from Form D. The resulting 13-item Analytical Reasoning

test has been designated Form H.[2]

To summarize thus far, it was decided that the following 13 items will be included on Form H, the new version of the Analytical Reasoning test: "Cow," "Fish," and "Air" from Form D and "Heat," "Cook," "Applause," "Chair," "Eat," "Bread," "Bag," "Cloth," "Furniture," and "Artist" from Form E. The configurations of these 13 items correspond to the 13 playing boards that were used with Forms D and E.

The next steps in the development of this new worksample were deciding on the item order and developing a stanine scoring system for the items. These are described below.

Item Order

On the Foundation's standard tests, including Forms D and F of Analytical Reasoning, the items generally are ordered according to difficulty. For the new Analytical Reasoning worksample, time to correct solution was used as the index of item difficulty. The median and mean response times for the 13 items on Form H are displayed in Table 2. The ranks of each item based on median and mean response times are also presented in the table. The order of difficulty, disregarding ties, is the same for the medians and means with the exception of three items: "Bread," "Furniture," and "Air," are ranked 9th, 10th, and 11th, respectively, on the basis of median response time and 10th, 11th, and 9th, respectively, on the basis of mean response time.

As indicated in the right-hand column of Table 2, the items for Form H have been arranged in approximate solution-time order, taking into consideration that the difficulty of the items (particularly the earlier items on Form D) may be affected to some extent by their original placement within Form D or E. In deciding on the item order, we also took into consideration the board configuration and content of the items. The items on Form H have been ordered in such a way that no more than two items of the same diagram type or similar content (e.g., items referring to food) will be administered consecutively.

---

[2]Another 13-item set resulting from these analyses was administered at the Boston test office during a study of software engineers. This version was designated Form G. Forms G and H consist of the same items with one exception: the "Breakfast" item is included on Form G as one of the 6-chip symmetric circular items (Diagram 4), whereas the "Fish" item is included on Form H as one of the items of this diagram type. The 13 items of Form G were presented in the following order: "Heat," "Eat," "Breakfast," "Chair," "Applause," "Cow," "Cook," "Bag," "Bread," "Furniture," "Air," "Artist," and "Cloth." This administration order corresponds to the presentation order of the 13 playing boards used with Forms D and E. The items on Form G were scored with the stanine scoring system used with the items on Form H (described in a later section of this report). The estimate of the alpha reliability of Form G is .841 using the logarithms of the item scores and .844 for the summed stanine scores. This reliability estimate for Form G is slightly inflated, however, because two of the items ("Breakfast" and "Cow") are almost identical.

Table 2

Difficulties of Items on Analytical Reasoning Worksample 244 H

Item #, Form D or E	1st chip	Median solution time	Rank based on mdn.	Mean solution time	Rank based on mean	Type of diagram[a]	Item #, new form
1 E	HEAT	.03	1	.04	1	1	1
8 E	EAT	.10	2	.14	2	3	2
7 E	CHAIR	.12	3	.16	3	2	3
5 E	APPLAUSE	.16	4	.22	4	4	5
2 D	COW	.19	5	.24	5	4	4
3 E	COOK	.24	7	.34	7	6	6
10 E	BAG	.24	7	.34	7	4	7
4 D	FISH	.24	7	.34	7	4	9
9 E	BREAD	.28	9	.41	10	7	11
12 E	FURNITURE	.29	10	.46	11	7	8
6 D	AIR	.31	11	.39	9	6	10
13 E	ARTIST	.42	12	.60	12	5	12
11 E	CLOTH	.57	13	.74	13	7	13

[a]Codes for the diagram types are as follows:

1 = 3-chip symmetric branched
2 = 5-chip symmetric branched
3 = 5-chip symmetric circular
4 = 6-chip symmetric circular
5 = 6-chip symmetric middle-converging
6 = another 6-chip symmetric middle-converging design distinct from Diagram 5
7 = 7-chip asymmetric circular

Scoring

In the analysis of the new Analytical Reasoning test to this point, continuous scores for each item were used. That is to say, the examinee's time to correct solution was recorded, and then a computer program generated the logarithm of that value in order to obtain an item score that approximated interval-level measurement. An examinee's total score was computed by adding together these logarithms of the item times. This scoring procedure is not practical with an individually administered test, however, because it is inconvenient to carry out by hand.

At present many of the Foundation's tests are scored using some type of time-point scale. For example, items on the standard Analytical Reasoning tests (Forms D and F) are scored based on a four-point scale that divides examinees into roughly equal-sized groups, such that about 25% of examinees receive one point, 25% receive two points, and so on. The difficulty with this type of scale is that it corresponds to a percentile rather than interval-level scale. As a result, more-recently developed Foundation tests, such as Number Facility, are scored using a stanine scale (Anastasi, 1976). As noted by Schroeder and Bethscheider (Statistical Bulletin 1988-6), "this method provides near-interval-level measurement at the item level while using a manageable number of whole-number scores" (p. 4).

In order to achieve approximately interval-level measurement for Form H, it was decided to score the 13 items using the traditional stanine method ($z = .5$) described in Statistical Bulletin 1988-6. That is to say, stanine scores for each item on the worksample were determined using time to correct solution as the basis for the nine time-point intervals, so that about 4% of the examinees received one point, 7% two points, 12% three points, 17% four points, 20% five points, 17% six points, 12% seven points, 7% eight points, and 4% nine points.

Two additional points regarding scoring can be made. First, Item 1 (the 3-chip "Heat" item) was so easy that 65% of the examinees arranged the chips correctly in .03 minute or less and over 95% arranged them correctly in .07 minute or less. This meant that it was not possible to separate examinees' scores on Item 1 into nine groups in such a way that the distribution fit a normal curve. Therefore, for Item 1, a 5-point scale based on approximately equal z intervals was used, with scores ranging from 1 to 5.

Second, examinees were told to stop working on each item after a specified amount of time had elapsed (e.g., at 1.5 minutes on the Form E items). Consequently, for several items on Form H, it was not possible to distinguish the lower points on the stanine scale. Approximately 21% of the examinees did not answer the "Cloth" item correctly in the allotted time. This group spanned the 1-point, 2-point, and most of the 3-point intervals. Similarly, on the "Artist" item, about 13% of the examinees did not arrange the chips correctly in the allotted time. This group spanned the 1-point, 2-point, and some of the 3-point intervals. As a result, an intermediate point value (namely, 2) was assigned to these groups. In other words, at this time, there is no 1-point value for either Item 12 or Item 13 of Form H. It is therefore recommended that examinees be allowed to work on the "Artist" and "Cloth" items for at least 2.00 minutes apiece, until we have enough data to distinguish the 1-point range for these two items.

Furthermore, almost 8% of the examinees did not answer the "Furniture" item correctly in the allotted time, 7% the "Fish" item, 6% the "Air" item, and 5% the "Bread" item. For these groups, which spanned the 1-point and part of the 2-point intervals, a point value of 1 was assigned. Again, it is recommended that examinees be allowed to continue work on these items beyond their original Form D or E termination times, until there is enough data to differentiate the full 2-point group from the 1-point group. (Specifically, the termination times for the "Fish" and "Air" items should be extended beyond 1.0 minutes, and the termination times for the "Furniture" and "Bread" should be extended slightly beyond 1.5 minutes.)

Based on the stanine scoring method described above, total scores on Form H can range from a minimum of 15 to a maximum of 113. Actual stanine scores on the full worksample for the examinees in this project ranged from 15 to 108, with a mean of 62.9 (*SD* = 14.0). With the exception of Item 1, each item of Form H had a mean stanine score of 5 and an *SD* of approximately 2, as should be the case. For Item 1, the mean "stanine" score was 3, with an *SD* of 1.

Form H has an alpha reliability of .831 for the summed stanine scores, compared with the reliability estimate of .832 obtained using the logarithms of the item scores. In other words, when stanine rather than continuous scores are used, there is a reduction of only .001 in the magnitude of the reliability coefficient.

Relationships of Item Characteristics to Item Quality and Item Difficulty

In order to obtain a better understanding of what distinguishes the more-effective Analytical Reasoning items from the less-effective items, the effects of various item characteristics on item quality and item difficulty were investigated. Specifically, the relationship of item features to item quality was explored by looking for trends among the item-total correlations of items with a shared feature, such as asymmetry or a specific type of diagram on the playing board. For this part of the study, then, interest in the item-total correlations focused on identifying which item features, if any, differentiate the better items from the poorer ones. For example, if one set of items with a shared feature (symmetry, for instance) tend to have higher item-total correlations, on average, than another set of items with a shared characteristic (asymmetry, for instance), then whether or not an item was symmetric might be an important consideration in terms of item quality. Similarly, as a secondary analysis, the item characteristics of the most-difficult and least-difficult items were also examined for trends, using mean time to correct solution as the index of difficulty. Note that all the item features studied involve some aspect of the item configurations and not the semantic content of the items. The findings for each of these configurational features (described earlier in the Method section) are discussed below.

1. Number of word-chips involved. In general, the fewer the number of chips involved, the easier the item, although several of the 5- and 6-chip items were among the most difficult of the Analytical Reasoning items. Regarding item quality, the average item-total correlation for the 3-chip items was .39; for the 5-, 6-, and 7-chip items, the average item-total correlations were .33, .40, and .41, respectively. It cannot be concluded, however, that 5-chip items are less-effective than items involving six or seven chips. Although two of the four 5-chip items have relatively low item-total correlations (i.e., .24 and .29), the other two 5-chip

items have acceptable item-total correlations that are comparable to those of the 6- and 7-chip items (i.e., .36 and .40). It appears, then, that number of word-chips is probably contributing to the difficulty but not the quality of the items.

2. Symmetric versus asymmetric configuration. Although there was a tendency for examinees to take longer to solve the asymmetric items than the symmetric items, this trend is confounded by another item characteristic--number of chips involved--because all the asymmetric items are 7-chip items and vice versa. Consequently, no conclusions can be reached regarding the relationship of asymmetry to item difficulty. The average item-total correlation for the symmetric items was .38 and for the asymmetric items, .41. These data provide little support for the notion that asymmetric configurations result in less-effective items than symmetric diagrams.

3. "Circular" versus "middle-converging" versus "branched" configuration. For the circular items, the average item-total correlation was .40; for the middle-converging items, .39; and for the branched items, .37. Furthermore, easier items were not distinguishable from more-difficult items on the basis of this item feature. In other words, this particular characteristic does not appear to be a contributing factor to either item quality or difficulty.

4. Specific type of diagram. Corrected item-total correlations for the eight items that can be described as 6-chip symmetric and circular (identified earlier as Diagram 4) ranged from .29 to .51, with an average item-total correlation of .41. For the six 7-chip asymmetric and circular items (Diagram 7), item-total correlations ranged from .35 to .46, also with an average correlation of .41. For the four 6-chip symmetric and middle-converging items (Diagram 6), item-total correlations ranged from .32 to .43, with an average correlations of .40. (Diagrams 1, 2, 3, and 5 were each administered only once during the standard worksample and once during the experimental worksample, so that examination of their correlations for trends was not particularly meaningful.) In general, no set of items with the same board configuration could be said to be better, on the average, than items of a different configuration.

Not surprisingly, Diagram 1--the only 3-chip configuration--was the easiest for examinees, and Diagram 7--the only 7-chip configuration--was the hardest. The two 5-chip configurations (Diagrams 2 and 3) tended to be of comparable difficulty to each other. Among the 6-chip configurations, the trend was for Diagram 5 to be more difficult than Diagrams 4 and 6.

5. Upside-down as well as right-side-up solution. The item-total correlations for the 16 items that have both right-side-up and upside-down solutions ranged from .29 to .51, with an average item-total correlation of .40. For the 10 items with only a right-side-up solution, the item-total correlations ranged from .24 to .43, with an average of .37. In other words, there was little difference in item quality between these two types of items. In addition, the most-difficult and least-difficult items could not be distinguished on the basis of this item feature.

In addition, a sixth item feature was investigated, namely, the number of chips that must be moved from one side of the board to the other in order to obtain the correct configuration. When comparing items that share the same board configuration, one might suppose that the more chips an examinee must move from one side of the board to the other in order to solve an item, the more time the examinee will take to place those chips in their correct positions

and, therefore, the more difficult the item would be. Moreover, it might seem that the number of required across-the-board movements should have no bearing on the effectiveness of an item. This description of the relationship of chip placement to item quality and difficulty is rather simplistic, however, and takes into account only the motor aspect of this item characteristic. More to the point is whether or not the effectiveness of the Analytical Reasoning items is influenced by the cognitive process associated with this feature. (This cognitive process is particularly relevant with regard to the asymmetric items because the three word-chips that go together must always be placed on the left-hand side of the diagram, while the two word-chips that go together must always be placed on the right-hand side. For the symmetric items, on the other hand, the word-chips that go together can be placed on either side of the diagram.)

By way of illustration, suppose that the two chips that go together on the right-hand side of the asymmetric diagram are presented on the left-hand side of the board. If the examinee initially places them on the left side where the three chips belong, he may then try to arrange the remaining chips around these two incorrectly placed chips before leaving this "blind alley" and correctly moving those two chips to the right-hand side of the diagram. Inasmuch as pursuing blind alleys precludes moving on to the correct placement of the word-chips and is not well-correlated with reasoning ability, then it might be expected that the more across-the-board chips movements an asymmetric item requires, the less-effective as well as more-difficult the item.

It is true that the 7-chip asymmetric circular items that entailed the movement of only one chip from one side of the board to the other were less difficult than those that involved the movement of three or four chips across the board. Nevertheless, the one item that required across-the-board movement of four chips (the "Broom" item) was less difficult than the two items that involved across-the-board movement of three chips (the "Country" and "Cigarette" items). These data, however, may be confounded by the fact that all the items that required across-the-board movement of three or more chips were from Form D, while all the items that involved across-the-board movement of only one chip were from Form E. That is to say, part of this difference in item difficulty may be related to practice effects.

Of greater interest, however, is the finding that the number of required across-the-board movements does not appear to differentiate the better asymmetric items from the poorer ones. The average item-total correlation for the items that involved across-the-board movement of only one chip was .42; for the items that required across-the-board movement of three or four chips, the average item-total correlation was .39. Consequently, this item feature does not appear to play a major role in influencing the effectiveness of the asymmetric items.

With regard to the 5- and 6-chip symmetric circular items, the easier items were not distinguishable from the more-difficult items on the basis of this item characteristic. Moreover, those items that required one across-the-board movement had an average item-total correlation of .42, and those that required two across-the-board movements had an average item-total correlation of .41, while those for which no across-the-board movements were required had an average item-total correlation of .40 (that is, .40 when the "Attack" item was excluded from the calculation, or .34 when the "Attack" item was included). In other words, there was no discernible pattern for the symmetric circular items--those that could be

solved without any across-the-board chip movements were no better (and no less difficult) than those requiring the movement of one or two chips to the other side of the board.

To summarize, it was found that, in general, the fewer the number of chips involved, the easier the item. Among the 6-chip items, those with the symmetric circular configuration (Diagram 4) and those with one of the symmetric middle-converging configurations (Diagram 6) tended to be less difficult than those with another symmetric middle-converging configuration (Diagram 5). No conclusion could be reached regarding the relationship of asymmetry to item difficulty because asymmetry is confounded with number of chips.

More important, no systematic relationships of item characteristics to item quality were identified. Some of the Analytical Reasoning items were better than others, independent of their configurational features. The same diagram, for instance, is involved in Items 2, 4, 5, and 10 on both Forms D and E, and yet their item-total correlations range from .29 to .51. This suggests that factors other than type of diagram play major roles in the effectiveness of these identically configured items. One consideration may be how widespread is the relevant knowledge needed to solve a specific item, such as the relationship between silkworms and stockings. Perhaps the larger issue, though, is the clarity of the conceptual structures represented in the items, as a function of both the clarity of the individual terms (e.g., *country* has multiple common meanings) and the relationships between the terms (e.g., the relationship between *animal* and *dog* is clearer, at least in terms of order, than the relationship between *attack* and *enemies*). That is to say, the clarity of the relationships among the words or concepts in all likelihood contributes more to item quality than the six item characteristics enumerated above.

Recommendations for Further Item Development

At this point we have accomplished the primary objective of this project, which was to construct a more reliable form of the Analytical Reasoning test, but the Foundation still needs to identify another pool of items that might serve in the future as replacements or additions to the standard test or as items for an alternate form of the test. If the Foundation is interested in replacing the least effective of the new test items, it should concentrate on developing replacements for the "Cloth," "Eat," and "Heat" items because these three items have correlations of less than .40 (but greater than .35) with total test score. It may, however, be more important for the Foundation to work on the development of an alternate form of the test, particularly if further reliability analyses confirm that the alpha coefficient of the new test falls in the acceptable range (i.e., is greater than .80).

Several of the items from Forms D and E that are not included on Form H have item-total correlations that might make them acceptable items on an alternate form, namely "Cake," "Human Being," and "Roast." ("Country" and "Breakfast" also have acceptable item-total correlations, but they should be excluded because the former is one of the "silkworm" items and the latter is too similar to the "Cow" item on Form H.) The other leftover items from Forms D and E would require some modification or refinement (e.g., making the relationships among the words or concepts clearer) before they could be considered for inclusion in the item pool. Other potential sources of items are previously developed group-administered or experimental forms of the test, such as Worksamples 696A and 696B, although such items would likely need to be revised to some extent before being used again

by the Foundation. Therefore, for the most part, new rather than revised items will probably need to be written.

In light of the aforementioned findings regarding the relationship of item characteristics to item quality, it is recommended that any further development of test items focus on the clarity issue. The construction of new items need not be limited to one type of diagram or even one type of item (such as symmetric or those with only right-side-up solutions), although it is suggested that more emphasis be placed on content areas that are not related to food. In addition, the most efficient way to construct an alternate form would be to develop items that utilize the seven board configurations used with the standard test. The Foundation's most pressing need would be for items that utilize a diagram for which few effective items currently exist. Above all, items for the following diagrams would be needed: (a) one 5-chip symmetric branched item (or a revised "Auto Factory" item); (b) one 5-chip symmetric circular item (or a revised "Attack" item); (c) two 6-chip symmetric circular items (one of which could be a revised "Cushion" item); (d) one 6-chip symmetric middle-converging item (Diagram 5); and (e) one 7-chip asymmetric circular item or a modified "Country" item without the word-chip *silkworm*.

Correlations and Factor Analysis

To assess the relationships among the Analytical Reasoning items and determine whether they are measuring a single underlying trait, correlational analyses and factor analyses were performed on the item data. First, zero-order correlations (simple Pearson product-moment coefficients) among the 26 standard and experimental items were computed in order to evaluate the magnitude of their relationships. Next, several exploratory factor analyses were performed. Estimates of the number of factors necessary to represent the item data were obtained from principal components analysis. Extracted factors were then rotated to the varimax criterion, and the resulting factor solutions assessed in terms of interpretability.

Examination of the pattern of correlations among the 26 items showed that the "Auto Factory," "Attack," and "Cushion" items correlated least with the other items. The range of correlations for the "Auto Factory" item with the other 25 items was .10 to .22, with only two correlations greater than .20; for the "Attack" item, the range was .10 to .21, with only one correlation greater than .20; and for the "Cushion" item, the range was .10 to .23, with only one correlation above .20. In contrast, the "Breakfast" item on Form E had moderate to high correlations (>.30) with 15 (60%) of the items. The "Applause" and "Chair" items each correlated moderately to highly with 9 (36%) of the items.

Correlational analyses also were employed as a means for exploring the amount of agreement among items with a shared characteristic, such as asymmetry or a specific type of diagram on the playing board. For example, when the correlations among items with the same board configuration were compared, the findings were as follows.

1. Correlations among the eight 6-chip symmetric and "circular" items (Diagram 4) ranged from .15 to .46. With the exception of the "Cushion" item, their highest correlations were with each other, although the "Breakfast" and "Applause" items also had moderate correlations with a number of items with other configurations.

2. Correlations among the six 7-chip asymmetric and circular items (Diagram 7) ranged from .19 to .35, with three of the 15 correlations greater than .30. In general, their highest correlations were with each other or with items of similar (i.e., circular), but less complex, 6-chip symmetric configurations.

3. Correlations among the four 6-chip symmetric and middle-converging items (Diagram 6) ranged from .20 to .27, but their highest correlations were with items with "circular" configurations.

(As mentioned previously, the other four diagram types were each administered only once during the standard worksample and once during the experimental worksample, so that examination of their correlations for trends was not particularly meaningful.)

In general, sets of items with the same board configuration cannot be said to be in more agreement with each other, on the average, than with items of different configurations. Likewise, no particular trend emerged for any of the other item characteristics mentioned earlier, including asymmetry and the number of required across-the-board chip moves. In other words, items that share a configurational feature do not appear to be more highly related to each other than to other items.

Initial principal components analysis with pairwise deletion extracted three factors with eigenvalues greater than unity, namely 6.89, 1.28, and 1.14. The first factor accounted for 26.5% of the total variance, while the second and third factors accounted for 4.9% and 4.4% of the total variance, respectively. None of the remaining factors accounted for more than 3.8% of the variance. From the scree plot it appeared that a one-factor model was the most appropriate for describing the relationships among the items. This indicates that the items are measuring a single underlying trait.

Although one factor appeared to be sufficient to represent the relationships among the Analytical Reasoning items, additional exploratory analyses were performed. Specifically, both the two-factor and the three-factor solutions were examined. Not surprisingly, inspection of the unrotated factor matrix showed Factor 1 to be a general factor in both cases, with all the items having their highest loading on that factor; the other factors were uninterpretable. Two-factor and three-factor rotated solutions also were examined in order to find out which items loaded on "non-general" factors, but again the second and third factors were uninterpretable. Thus, distinct factors (beyond the general factor) did not emerge for any of the item characteristics under study.

Sex and Age Effects

Sex and age differences in performance on Analytical Reasoning were evaluated using the regression approach to analysis of variance (ANOVA) of the test scores by sex and age. For these analyses, the sample was partitioned into the following ten age groups: (a) 14-16 years of age, (b) 17-18, (c) 19-20, (d) 21-23, (e) 24-26, (f) 27-30, (g) 31-34, (h) 35-39, (i) 40-47, and (j) 48 and older.

The ANOVAs for transformed time scores on the standard and experimental worksamples yielded similar results. For Forms D, E, and F, there was neither a significant

sex difference nor a significant age-by-sex interaction. There were, however, significant age effects for these worksamples: for Form D, $F(9;1,473) = 3.78$, $p < .001$; for Form E, $F(9;1,473) = 4.67$, $p < .001$; and for Form F, $F(9;1,473) = 3.88$, $p < .001$. When scores are adjusted for sex, the proportion of variance that is accounted for by age alone is small ($\omega^2 = .02$ for all three forms). This is comparable to the effect of a correlation of .14.

Likewise, the ANOVA for stanine scores on Form H yielded a significant main effect for age, $F(9;1,482) = 5.11$, $p < .001$. After controlling for sex, the effect of age accounted for 2% of the variance in test scores ($\omega^2 = .02$), which is equivalent to the effect of a correlation of .14. In Cohen's (1988, pp. 24-27) terms, there are relatively small differences among the age groups. Males and females did not differ significantly in their performance on the new worksample. The age-by-sex interaction also was not significant.

The age curve for Form H is graphically depicted in Figure 3. The plotted values are the mean scores (in standard deviation units) of the 10 age groups, adjusted for the effects of sex. The extent of the adult plateau is difficult to establish precisely with the current sample, although the adult plateau appears to be in evidence by age 14 and extends through the late twenties. Performance declines modestly from the thirties through the mid-forties, with a sharper decline observed after age 47.

Additional Analyses

Besides the series of analyses discussed above, several other analyses were performed on the Analytical Reasoning data. These include (a) calculation of additional reliability estimates for Form H as well as Forms D and F, including reliability estimates controlled for the effects of age and sex and reliability estimates for two subsamples of examinees--the adult plateau and a male subsample; (b) comparison of the reliability of Form D with Forms H and CM; and (c) reliability analyses for an alternate form composed of items from Forms D and E that are not included on Form H. Also explored were (a) the relationship between Forms F and H; (b) the relationship between Form G and Forms D, F, and H; and (c) the relationship of sex and age to performance on the "silkworm" items. The results of these analyses are reported in this section.

Additional Reliability Analyses for Form H

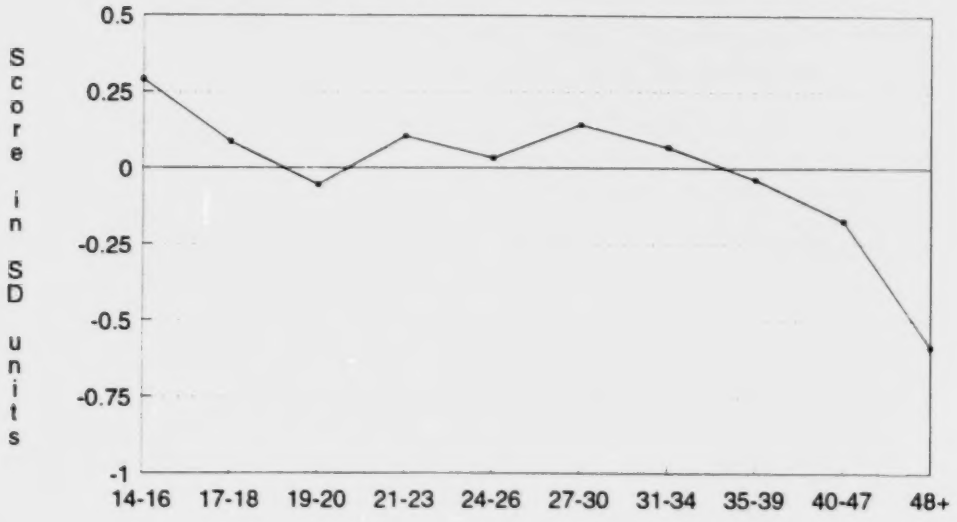
The alpha coefficient of .831 reported earlier in this document for the new Analytical Reasoning worksample (based on the summed stanine scores) is somewhat inflated because the effects of age and sex were not removed from the analysis. In order to estimate approximately how much the reliability of Form H would drop if sex and age were removed from the analysis, the following analysis was performed.[3] First, the items of the test were arranged temporally according to their original order of administration (i.e., the items from Form D before the items from Form E); those items in the odd-numbered positions were assigned to one half of the test and those items in the even-numbered positions to the other

[3]It is difficult to correct for sex and age explicitly in an alpha coefficient because that would require correcting scores on individual items, which tends to be an unreliable procedure.

half. Then two estimates of the test's split-half reliability were calculated--one corrected for the effects of age and sex and the other with no control for age and sex. The uncorrected reliability was computed as the split-half correlation between the scores for the two halves of the test, which was then adjusted by the Spearman-Brown correction, yielding a reliability estimate for the complete worksample of .845 when the effects of age and sex are not controlled. To adjust for the effects of age and sex, the score for each half was regressed separately on sex, age, age-squared, and age-cubed. The residuals from these regressions represent the scores for the two halves corrected for age and sex. The split-half correlation between these residuals was computed and then adjusted by the Spearman-Brown correction, yielding a corrected reliability estimate of .841. The difference between these two split-half reliability estimates (i.e., .004) is an approximation of the loss in Form H's reliability if the effects of age and sex were removed. In other words, the alpha reliability of Form H would probably be about .827 if the effects of age and sex were removed.

Figure 3

Age Curve for Analytical Reasoning Worksample 244 H



Note. Plotted values for Form H are the summed item stanine scores (in standard deviation units), adjusted for the effects of sex. Stanine scores for each item are based on a traditional stanine scale (z = .5) with time to correct solution used as the basis for the nine time-point intervals, so that about 4% of the examinees received one point, 7% two points, 12% three points, 17% four points, 20% five points, 17% six points, 12% seven points, 7% eight points, and 4% nine points. Zero point of scale used in graph corresponds to a stanine score of 62.9, with each *SD* unit equivalent to 14.0 stanine points. Plotted values are not smoothed.

An alternative method for estimating loss in a test's reliability due to age and sex effects is to conduct reliability analyses on homogeneous segments of the examinee sample. Because scores on Analytical Reasoning are related significantly to age but not sex, for these analyses we were more interested in controlling for age effects than sex effects. Therefore, alpha coefficients were computed for the following subsamples: (a) the adult plateau, restricted for this analysis to examinees aged 22 to 28; and (b) males aged 19 or 20, a highly homogeneous group for which reliability estimates have been calculated by the Foundation in the past. For the adult plateau, an alpha coefficient of .829 was obtained, for a loss in reliability of .002 relative to the uncorrected coefficient. For the 19- and 20-year-old males, an alpha coefficient of .833 was obtained, for a slight gain in reliability of .002. In both analyses, the samples sizes were reasonably adequate for estimating alpha reliabilities (*n*s = 336 for the subsample of adults and 101 for the subsample of males).

Additional Reliability Analyses for Forms D and F

Using the Foundation's current 1-to-4-point scoring system, alphas of .746 and .703 were obtained for Forms D and F, respectively (compared with .678 and .612, respectively, using untransformed item times and .759 and .724, respectively, using logarithms of the item times). When the effects of age and sex were controlled using the procedures described earlier, the alpha reliability is estimated to be .744 for Form D and .700 for Form F. Additional reliability estimates using the 4-point scoring system were calculated for the two subsamples of examinees studied earlier. The resulting alphas for Forms D and F were .733 and .691, respectively, for the adult plateau and .781 and .728, respectively, for the 19- and 20-year-old-male subsample.

The time intervals for the 4-point scoring scale have not been reviewed for a number of years. Therefore, the time-point intervals for the items were revised temporarily to more evenly distribute the item scores across the quartiles. (This updating primarily affected the scoring of the "Clothes," "Auto Factory," "Cake," "Broom," "Cushion," and "Attack" items and, to a lesser extent, the "Human Being" item; especially, the 3- and 4-point time intervals for the "Clothes" item are too strict.) Coefficient alphas for Forms D and F were then recalculated using this updated scoring system. For Form D, an alpha of .750 was obtained for the full sample and for Form F, an alpha of .709, indicating that very slight gains in reliability might be achieved by revising the current quartiles; nevertheless, the reliabilities of Forms D and F based on an updated 4-point scoring system would continue to be lower than is desirable by Foundation standards. Reliability estimates using the updated 4-point scoring system were also calculated for the two subsamples: for the adult plateau, alphas of .736 and .697 were obtained for Forms D and F, respectively; for the 19- and 20-year-old males, alphas of .780 and .727, respectively.

As mentioned previously, the Foundation's 1-to-4-point scoring system corresponds to a percentile rather than interval-level scale. To evaluate the effects of scoring the standard Analytical Reasoning test using a stanine scale, time-point intervals based on the traditional stanine method were determined for the items on Forms D and F, and the items were then rescored. When stanine scores were used in the calculation of alpha, a reliability estimate of .771 resulted for Form D and a reliability estimate of .734 for Form F. Although the magnitude of the reliability coefficients for Forms D and F would be increased somewhat by using stanine scores rather than the current 1-to-4-point scores, the reliability of these forms would still fall short of the Foundation's recommended minimum standard of .80.

Comparison of Form D with Form H. In order to compare the 13-item standard worksample with the new 13-item worksample on the basis of their reliabilities, both sets of items should be scored using the same type of scoring system. For this analysis, the test items on both forms were scored using the traditional stanine method. As reported above, when stanine scores were used in the calculation of alpha, a reliability of .771 was obtained for Form D, compared with a reliability of .831 for Form H, for a difference in magnitude of .060.

There are two caveats regarding this result. First, the reported reliabilities for these worksamples could be slightly higher (or lower) than their true reliabilities due to chance alone, although these effects are probably very small because the sample size in this study is quite substantial. Second, because Form H is composed predominantly of items from Form E, which was always administered after Form D, the reported reliability of Form H may be inflated somewhat due to practice effects. That is to say, although it is likely that the reliability of Form H is higher than the reliability of Form D, the magnitude of the difference between their reliabilities may actually be smaller than is reported here.

Comparison of Form D with Form CM. As noted in the Introduction, before this project the most recent internal-consistency reliability estimate for Analytical Reasoning was .65. This reliability estimate was computed as the odd-even correlation for Form CM (based on males aged 19 and 20) adjusted by the Spearman-Brown correction. Examinees who did not complete the worksample were excluded from the analysis (about five percent of the sample). When this same method of calculation was used with the 13 items of Form D in the present study, a reliability estimate of .76 was obtained.

Reliability Analyses for an "Alternate" Form

Earlier in this report, the recommendation was made that the Foundation develop an alternate form of the Analytical Reasoning test. To determine whether a combination of items from Forms D and E that are not included on Form H might adequately serve as an alternate test, the psychometric properties of a test composed of 11 of the 13 leftover items were assessed. Specifically, the following items were chosen for inclusion on this trial alternate test: "Roast," "Human Being," "Cake," "Auto Factory," "Attack," "Broom," "Cushion," "Country," "Cigarette," "Arrow," and "Astronaut." One of the "silkworm" items--"Clothes"--was omitted from the analyses along with the "Breakfast" item, which is too similar to the "Cow" item on Form H. (The other "silkworm" item--"Country"--was included in the analyses because it showed a good item-total correlation even though it involves acquired knowledge.)

The corrected item-total correlation for each item on the alternate test was determined by correlating the item score with the total score on Form F (not the total score for the alternate test), adjusted for part-whole overlap (refer to Table 1). Item-total correlations for the 11 items ranged from .26 to .46. The "Attack, "Auto Factory," and "Cushion" items displayed relatively low item-total correlations. The remaining eight items correlated at least .31 with the score for Form F.

When all 11 items were included in the total score, an alpha of .711 resulted, which is lower than is desirable by Foundation standards. All 11 items contributed positively to the internal consistency of the alternate test. When the "Attack" item was excluded from the calculation, a reliability of .705 was obtained; alternatively, excluding the "Auto Factory" item

yielded a .702 coefficient; when both were excluded, a reliability of .697 resulted. When all three items with corrected item-total correlations below the level of acceptability were deleted from the total score, the reliability estimate was reduced approximately .027; that is, alpha equaled .684 for the 8-item alternate test. Because the "Country" item depends in part on acquired knowledge about silkworms and will not be used again by the Foundation without some modification, the reliability estimate of the alternate test was recalculated with this "silkworm" item excluded from the analysis, resulting in an alpha of .637 for the 7-item alternate test.

An estimate of the reliability of a 13-item alternate test was computed by applying the Spearman-Brown correction to the 7-item test's reliability, yielding a reliability estimate of .765 for a full-length alternate test. This suggests that it might be possible to construct an alternate form with close-to-acceptable reliability by using seven of the leftover items from Forms D and E (namely, "Roast," "Human Being," "Cake," "Broom," "Cigarette," "Arrow," and "Astronaut") and modifying or replacing the "Auto Factory," "Attack," "Cushion," "Country," "Clothes," and "Breakfast" items.

Reliability estimates using stanine scores were also calculated for two subsamples of examinees. When all 11 items were included in the total score, an alpha coefficient of .713 was obtained for the adult plateau, and an alpha of .727 resulted for the 19- and 20-year-old males.

The correlation between the 11-item alternate form and Form H, uncorrected for attenuation, is .74 when summed stanine scores are used. The disattenuated correlation between the two forms is .97.

Other analyses

Relationship between Forms F and H. In examining the relationship between the current standard test and the new test of Analytical Reasoning, the question that needs to be addressed is: Have the proposed changes to the test altered the underlying construct in any way? If scores on Forms F and H are highly related to each other, then it may be said that the two tests measure the same construct.

To determine the magnitude of the relationship between the two forms, the disattenuated correlation between Forms F and H was calculated. Because the calculation of this correlation presumes that the items of the two tests are mutually exclusive, part of the gain derived by correcting the correlation for attenuation may be unjustified if the two tests share some items. In particular, because Forms F and H have three items in common, about 25% of the difference between their disattenuated correlation and their uncorrected correlation may be said to be unwarranted. Therefore, 25% of this difference was subtracted from the disattenuated correlation in order to produce a more accurate estimate of the corrected correlation between Forms F and H.

The correlation between Forms F and H, uncorrected for attenuation, is .78. The disattenuated correlation between the two tests approaches unity. Adjusting for the 3-item overlap using the procedure described in the preceding paragraph resulted in an estimate of .94 as the corrected correlation between Forms F and H. This substantial relationship

between the current standard test and the new test indicates that the proposed changes have not altered the construct underlying analytical reasoning.

Relationship between Form G and Forms D, F, and H. As noted earlier, Forms G and H are similar, except that the administration order of the test items is different, and Form G includes two nearly identical items ("Cow" and "Breakfast") whereas Form H includes the "Fish" item instead of the "Breakfast" item. As noted earlier, this means that the alpha reliability of Form G is inflated because of the repetition of the nearly identical items; independent of that consideration, the reliability of Form G is likely to be similar to the reliability of Form H.

Not surprisingly, the disattenuated correlation between the two forms approaches unity when item overlap is adjusted using the procedure described earlier. The disattenuated correlation between Forms G and D, which have two items in common, is .91; the disattenuated correlation between Forms G and F, which likewise share two items, is .92.

If the effects of age and sex were controlled, the magnitude of the reliability coefficient for Form G would be reduced by approximately .002, which means that the alpha reliability of Form G would probably be about .842. Additional reliability estimates were calculated for two subsamples of examinees. An uncorrected alpha coefficient of .846 was obtained for both the adult plateau and the 19- and 20-year-old males, compared with a reliability estimate for .844 for the complete sample.

Relationship of sex and age to "silkworm" items. As noted earlier, the "Country" and "Clothes" items were removed recently from the standard Analytical Reasoning test because their solutions depend in part on acquired knowledge, such as the connections between "silkworm" and "stockings" and "silkworm" and "thread." It was hypothesized that younger examinees may not have this knowledge and therefore would not perform as well as older examinees, particularly females, on the two "silkworm" items. Therefore, sex and age differences in performance on the "silkworm" items were investigated using the regression approach to ANOVA of the item stanine scores by sex and age with scores on Form F as a covariate. For these analyses, the younger examinee group consisted of those aged 14 to 18, and the older group was composed of examinees 40 years of age and older.

With regard to the "Country" item, the ANOVA showed no significant main effect for sex or age. With regard to the "Clothes" item, the ANOVA showed a significant main effect for sex, $F(1, 480) = 11.10$, $p = .001$, with males outperforming females, but there was not a significant effect for age. The sex-by-age interaction was not significant for either item. These findings do not support for our hypothesis regarding the relationships of sex and age to performance on the two items that depend on acquired knowledge about silkworms.

GENERAL SUMMARY AND CONCLUSIONS

The purpose of this research project was to develop a more-reliable form of Analytical Reasoning using items from Worksample 244, Forms D and E. In addition, the Foundation had an interest in identifying which item characteristics, if any, contribute to the effectiveness, or quality, of the Analytical Reasoning items.

Factor analysis of the data indicated that all 26 items are measuring a single underlying trait. Findings from the internal-structure analysis of the items showed that the "Cigarette" and "Clothes" items were the most difficult for examinees. All but three items ("Attack," "Auto Factory," and "Cushion") demonstrated at least adequate item-total correlations (i.e., above .30) with Form F, the current standard form of the test. Two of the items ("Attack" and "Auto Factory") made negligible contributions to the internal consistency of the overall test when all the items were included in the total score; the remaining items contributed positively to overall test reliability.

In selecting items for the new, more-reliable version of the Analytical Reasoning test, several indices of item quality were employed. Initially, items were evaluated in terms of their corrected item-total correlations and/or unique contributions to the overall reliability of the test, and the alpha reliabilities for various combinations of the better items were calculated and compared. Obtained reliability estimates for some of the initial item combinations ranged from .829 (for a 13-item set) to .864 (for a 20-item set).

Based on the results of those analyses, two decisions regarding the item composition of the new test were then formulated. One was to restrict the number of items that could be selected for the new test to no more than 13 to 15, so that the new test would not take much longer to administer than the original 13-item standard test. The other was to include a 3-chip item on the test, even if it would otherwise be excluded on the basis of the selection criteria. The imposition of these conditions meant a sacrifice in reliability of at most .036, relative to the highest possible alpha, .864 for the 20-item set. Hence alpha would remain in an acceptable range (i.e., above .80).

At this point in the process of developing a more-reliable test, another selection criterion was imposed. This one required the selection of sets of items that made use of the 13 existing Analytical Reasoning playing boards without sacrificing reliability. Obtained reliability estimates for some of these item combinations ranged from .829 (for a 13-item set) to .842 (for a 15-item set).

On the basis of these findings, several decisions were made. The first was that the 13 playing boards from Forms D and E would in fact be utilized with the new test because the effect on the internal consistency of the new test would be minimal. The second was that a 13-item test would be developed rather than a 14- or 15-item test, given that all three were very similar in terms of overall test reliability. The decision to construct the new worksample from the most reliable of the 13-item combinations that utilized the 13 playing boards from Forms D and E resulted in the selection of the following set of items for the next standard form of the Analytical Reasoning test, which has been designated Form H: "Cow," "Fish," and "Air" from Form D and "Heat," "Cook," "Applause," "Chair," "Eat," "Bread," "Bag," "Cloth," "Furniture," and "Artist" from Form E.

Once the items for the new form were selected, the next steps in the development of this worksample were deciding on the item order and developing a stanine scoring system for the items. With regard to item order, the 13 items were arranged in approximate order of difficulty, bearing in mind that the difficulty of the items (particularly the earlier items on Form D) may have been affected to some extent by their original placement within Forms D and E. Also taken into account were the board configuration and content of the items, so

that no more than two items of the same diagram type or similar content would be administered consecutively. The item order chosen for the 13 items on the new worksample was as follows: "Heat," "Eat," "Chair," "Cow," "Applause," "Cook," "Bag," "Furniture," "Fish," "Air," "Bread," "Artist," and "Cloth."

In order to achieve approximately interval-level measurement for Form H, the items were scored using the traditional stanine method ($z = .5$), with time to correct solution as the basis for the nine time-point intervals. Because it was not possible to distinguish the lower points on the stanine scale for several items on Form H, it is recommended that, in a future study, examinees be allowed to work on these items beyond their original Form D or E termination times, until enough data are collected to distinguish the 1-point range for these items. Specifically, the termination times should be extended for the "Furniture," "Fish," "Air," "Bread," "Artist," and "Cloth" items. It is estimated that the alpha reliability of Form H (based on the summed stanine scores) would be about .827 if the effects of age and sex were removed.

The correlation between the current standard test and the new test, corrected for attenuation and adjusted for item overlap, is estimated to be .94. This substantial relationship between Forms F and H indicates that the proposed changes to the Analytical Reasoning test will not alter the underlying construct in any way.

In addition to developing a form of Analytical Reasoning with improved reliability, the Foundation was interested in obtaining a better understanding of what distinguishes the more-effective Analytical Reasoning items from the less-effective items in order to make recommendations regarding the composition of future items. Therefore, the effects of various item characteristics on item quality and item difficulty were investigated. The six item features studied were: (a) number of word-chips involved; (b) symmetry or asymmetry of the item configuration; (c) "circular" versus "middle-converging" versus "branched" configuration; (d) specific type of diagram; (e) right-side-up solution only versus upside-down as well as right-side-up solution; and (f) number of required across-the-board chip movements.

In terms of differentiating the easier items from the more-difficult ones on the basis of their item characteristics, it was found that, in general, the fewer the number of chips involved, the easier the item. No conclusion could be reached regarding the relationship of asymmetry to item difficulty because asymmetric items tend to have greater numbers of chips.

More important, no systematic relationships of item characteristics to item quality were identified. Of particular interest were two findings regarding the asymmetric items. The first was that asymmetric configurations do not appear to result in less-effective items than symmetric diagrams. The second was that number of required across-the-board movements does not appear to differentiate the better asymmetric items from the poorer ones. It was concluded that some of the Analytical Reasoning items are more effective than others, independent of their configurational features. One consideration may be whether acquired knowledge is needed to solve a specific item, although the larger issue may be the clarity of the conceptual structures represented in the items, as a function of both the clarity of the individual terms and the relationships between the terms. That is to say, the clarity of the

relationships among the words or concepts in all likelihood contributes more to item quality than the six item characteristics that we studied.

In light of the findings, several recommendations can be made regarding the direction of further item development on the Analytical Reasoning test. One recommendation is that any further development of test items focus on making the relationships among the words or concepts clearer. Moreover, it is recommended that future Foundation research be directed towards identifying acceptable items for an alternate form of the Analytical Reasoning test rather than refining the new standard form.

To determine whether a combination of items from Forms D and E that were not included on Form H might adequately serve as an alternate test, the psychometric properties of a test composed of 11 of the 13 leftover items were assessed. It was concluded that it might be possible to construct an alternate form with close-to-acceptable reliability by using seven of these leftover items and modifying or replacing the other items. The Foundation's most pressing need, then, would be for items that utilize diagrams for which insufficient items currently exist. Specifically, items for the following diagrams would be needed (presuming that the alternate test would consist of at least 13 items that use the same 13 playing boards as Form H): (a) one 5-chip symmetric branched item or a revised "Auto Factory" item (Diagram 2); (b) one 5-chip symmetric circular item or a revised "Attack" item (Diagram 3); (c) two 6-chip symmetric circular items, one of which could be a revised "Cushion" item (Diagram 4); (d) one 6-chip symmetric middle-converging item (Diagram 5); and (e) one 7-chip asymmetric circular item or a modified "Country" item without the word-chip *silkworm* (Diagram 7).

In summary, a new standard form of the Foundation's Analytical Reasoning test has been developed using items from Forms D and E. This new worksample, designated Form H, has an estimated alpha reliability of .827. Form H was constructed from the most reliable of the 13-item combinations that utilized the existing playing boards. The items selected for Form H, listed in order of presentation, are as follows: "Heat," "Eat," "Chair," "Cow," "Applause," "Cook," "Bag," "Furniture," "Fish," "Air," "Bread," "Artist," and "Cloth."

Now that a standard form of Analytical Reasoning with improved reliability has been developed, it is recommended that the Foundation conduct further research on the development of an alternate form of the test. In particular, a relatively reliable alternate form might be possible to construct by using some of the items from Forms D and E that are not included on Form H and then developing additional items.

REFERENCES

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: MacMillan.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Norusis, M. J. (1990a). *SPSS/PC+ 4.0 base manual.* Chicago: SPSS.

Norusis, M. J. (1990b). *SPSS/PC+ Statistics 4.0.* Chicago: SPSS.

Scoring guide for group worksamples and standard administration of individual worksamples. (1987). New York: Johnson O'Connor Research Foundation.

Statistical Bulletin 1974-13. *Reliability of Worksample 244CM, Analytical Reasoning.* M. Daniel. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 1988-2. *JOCRF test reliabilities and interpretation of test scores.* D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 1988-6. *Development of Wks. 436 IA.* D. H. Schroeder & J. K. Bethscheider. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 1989-2. *Construction of an alternative form of Analytical Reasoning - Wks. 244 E.* R. F. Kyle. Chicago: Johnson O'Connor Research Foundation.

Test Information Bulletin 1991-9. *Deleting the "silkworm" items from Analytical Reasoning, resulting in Wks. 244 FA.* J. K. Bethscheider & D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

BEST COPY AVAILABLE